Tutorial

MDM Medical Decision Making

Medical Decision Making
1–13

DOI: 10.1177/0272989X211026292
journals.sagepub.com/home/mdm

Simulating Study Data to Support Expected Value of Sample Information Calculations: A Tutorial

Anna Heath, Mark Strong, David Glynn, Natalia Kunst, Nicky J. Welton, and Jeremy D. Goldhaber-Fiebert

The expected value of sample information (EVSI) can be used to prioritize avenues for future research and design studies that support medical decision making and offer value for money spent. EVSI is calculated based on 3 key elements. Two of these, a probabilistic model-based economic evaluation and updating model uncertainty based on simulated data, have been frequently discussed in the literature. By contrast, the third element, simulating data from the proposed studies, has received little attention. This tutorial contributes to bridging this gap by providing a step-by-step guide to simulating study data for EVSI calculations. We discuss a general-purpose algorithm for simulating data and demonstrate its use to simulate 3 different outcome types. We then discuss how to induce correlations in the generated data, how to adjust for common issues in study implementation such as missingness and censoring, and how individual patient data from previous studies can be leveraged to undertake EVSI calculations. For all examples, we provide comprehensive code written in the `R` language and, where possible, Excel spreadsheets in the supplementary materials. This tutorial facilitates practical EVSI calculations and allows EVSI to be used to prioritize research and design studies.

Keywords
expected value of sample information, `R` tutorial, research design methods, simulation methods, value of information

Date received: December 21, 2020; accepted: May 20, 2021

Introduction

What Is EVSI and Why Is It Not Used More Frequently?

The expected value of sample information (EVSI) measures the value of reducing decision uncertainty by undertaking a proposed study with a given design.[1] Specifically, EVSI is the expected economic benefit of a study that collects additional information that aims to reduce uncertainty *before* making a decision.[2] In medical decision making, EVSI can be applied to a wide range of study designs, including clinical trials, to inform the relative effectiveness of treatments or observational studies to estimate baseline event rates. The expected net benefit of sampling (ENBS) is defined as the costs of a study subtracted from its (population-level) EVSI. Studies with high ENBS efficiently trade off information value and data collection cost. ENBS can then be used to optimize study design and prioritize research investments that offer value for money.[3,4] EVSI and ENBS can also support reimbursement decision makers as small values for EVSI and ENBS indicate that treatment recommendations should be made using existing evidence, rather than recommending the collection of further evidence before making a treatment recommendation. Despite these benefits of EVSI and ENBS, their practical application has been restricted by the difficulty of the computations required and by the small number of analysts who are familiar with its use.[5]

Corresponding Author
Anna Heath, Child Health Evaluative Sciences, Peter Gilgan Centre for Research and Learning, The Hospital for Sick Children, 686 Bay St, Fl 11, L4 East, Toronto, ON M5G 0A4, Canada.
(anna.heath@sickkids.ca)

How Is EVSI Computed?

In model-based health economic evaluations, EVSI is usually calculated using a simulation-based approach based on 3 main elements, each of which can increase the barrier to its implementation.[6] First, the model-based economic evaluation must be fully *probabilistic* (i.e., all relevant quantities must be parameterized and their uncertainty accurately characterized and encoded in probability distributions). In this setting, the optimum decision option is the one that maximizes *expected* net benefit, where expectation is taken over the parameter uncertainty.[1] Second, we must simulate plausible values for the data that would be collected in the proposed future study.[6] Third, we must update our parameter uncertainty using the simulated plausible study data from the previous step, potentially changing the optimum decision option.[7] This final step has traditionally been highly computationally demanding because it requires a large number of simulations.

The first and third elements of the process have been widely discussed. First, methods for developing probabilistic decision-analytic models are well established, since probabilistic analyses (PAs), also known as probabilistic sensitivity analyses, are required as part of health technology assessment (HTA) processes in many health systems.[8–12] Good practice guidelines and textbooks also guide the development of probabilistic decision-analytic models using evidence from the literature.[1,13–15] The third element has been facilitated by recently developed efficient approximation methods that have overcome the computational challenge of calculating EVSI using the simulated study data.[16–19] These approximation methods have recently been compared and evaluated.[20,21]

What Does This Tutorial Discuss?

This tutorial addresses the crucial second element, simulating plausible study data, which has not received sufficient attention in the literature to allow analysts to easily compute EVSI. Fortunately, simulating study data is a common task outside of HTA.[22,23] This tutorial highlights how these approaches[23–29] can be used to compute EVSI. We will present methods to simulate data using correlated and uncorrelated parametric distributions that incorporate *real-world* study challenges, such as loss to follow-up, and using a nonparametric approach with individual patient data (IPD) from previous studies. We aim to support the generation of realistic study data to improve the accuracy of EVSI calculations.[6] Coupled with the recent advancements in EVSI computation, this tutorial will facilitate the use of EVSI in practice to guide research prioritization and study design.

Background and Notation

This section provides a brief introduction to EVSI and the notation used throughout this tutorial. A more complete introduction to EVSI is included in other sources.[1,7,21]

Model-Based Decision Analysis

We are aiming to decide between a set of $d = 1, \ldots, \mathcal{D}$ interventions. We have a decision-analytic model that estimates the net benefit for each option d, given a vector of $\mathcal{P}$ input parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{\mathcal{P}})$. We consider that the model is a function that maps inputs $\boldsymbol{\theta}$ to strategy-specific net benefits NB_d, denoted $\mathrm{NB}_d(\boldsymbol{\theta})$. The inputs $\boldsymbol{\theta}$ represent *real-world* quantities (e.g., costs, relative treatment effects, disease progression on standard care, utilities, and disease prevalence), which are not known with certainty. Through a PA, we represent knowledge about these quantities via the joint probability distribution $p(\boldsymbol{\theta})$, which can be considered as describing the joint prior distribution for $\boldsymbol{\theta}$. The expected net benefit of the optimum decision given current knowledge is $\max_d \mathbb{E}_{\boldsymbol{\theta}}\{\mathrm{NB}_d(\boldsymbol{\theta})\}$. This expectation is usually estimated using Monte Carlo simulation (i.e., values of $\boldsymbol{\theta}$ are

Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON, Canada (AH); Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada (AH); Department of Statistical Science, University College London, London, UK (AH); School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK (MS); Centre for Health Economics, University of York, York, UK (DG); Harvard Medical School & Harvard Pilgrim Health Care Institute, Harvard University, Boston, MA (NK); School of Social and Community Medicine, University of Bristol, Bristol, UK (NJW); and Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA (JDGF). The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: AH was funded in part by an Innovative Clinical Trials Multi-year Grant from the Canadian Institutes of Health Research (funding reference number MYG-151207; 2017–2020), as part of the Strategy for Patient-Oriented Research. MS has no funding to declare. DG has no funding to declare. NK reports funding from the Research Council of Norway (276146 and 304034) and Link Medical Research during the conduct of the study and personal fees from Thermo Fisher Scientific outside the submitted work. NJW was supported by the NIHR Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol. JDGF was funded in part by a grant from Stanford's Precision Health and Integrated Diagnostics Center (PHIND). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

sampled from $p(\boldsymbol{\theta})$ and used to compute the average net benefit for each d) because it is usually not available analytically.

The Expected Value of Sample Information

Data to update information in $\boldsymbol{\theta}$ have value if they might change the optimum treatment. If we were to collect new data $\boldsymbol{x}$ and update our knowledge about $\boldsymbol{\theta}$ and the net benefits, the optimal decision would be the option that maximizes the expected net benefit, $\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{x}}\{\mathrm{NB}_d(\boldsymbol{\theta})\}$, conditional on the new data. However, before conducting a study, the data have not been collected, and so we compute the *expected value* of collecting additional data, where the expectation is taken with respect to the distribution of all plausible realizations of the data that the proposed study may generate. Thus, the data from the proposed study are a random variable, denoted $\boldsymbol{X}$, and are not yet observed. The expected value of the net benefit for the optimal decision given new information, averaged over the distribution of all possible datasets, $p(\boldsymbol{X})$, is $\mathbb{E}_{\boldsymbol{X}}[\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\{\mathrm{NB}_d(\boldsymbol{\theta})\}]$, and EVSI is the difference between this quantity and the expected net benefit under current information,

$$\mathrm{EVSI} = \mathbb{E}_{\boldsymbol{X}}[\max_d \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\{\mathrm{NB}_d(\boldsymbol{\theta})\}] - \max_d \mathbb{E}_{\boldsymbol{\theta}}\{\mathrm{NB}_d(\boldsymbol{\theta})\}. \quad (1)$$

The first and second terms in this equation are usually not available in closed form and must be estimated using simulation methods.

$\boldsymbol{X}$ is the complete set of quantities that would be collected during the study. In reality, this dataset may include mismeasured quantities, missing values, and measurements taken at times that deviate from the study design, which should be reflected in our distribution for $\boldsymbol{X}$.[6] Furthermore, a model parameter could be informed by different study designs (e.g., relative effectiveness can be estimated through a randomized controlled trial or through an observational study using suitable methods, which would result in different $\boldsymbol{X}$).

Efficient Methods for Computing EVSI

The "standard" approach to EVSI estimation uses a nested Monte Carlo scheme that requires a large number of samples from the posterior distribution of the model parameters given sampled data, $p(\boldsymbol{\theta}|\boldsymbol{x})$, (an "inner loop") nested within an "outer loop" that samples a large number of simulated datasets $\boldsymbol{x} \sim p(\boldsymbol{X})$. If the numbers of inner-loop and outer-loop samples are $\mathcal{N}_i$ and $\mathcal{N}_o$, respectively, the decision-analytic model must be evaluated $\mathcal{N}_i \times \mathcal{N}_o$, requiring days or even months to complete the required computation.[17] However, recent methods for computing EVSI decrease this time to seconds via approximations that either reduce $\mathcal{N}_o$, the number of simulated datasets required, or avoid the inner loop altogether.[16–20]

Approaches to Simulating Study Datasets

We now discuss how to simulate plausible study datasets. For some EVSI computation methods, only a *summary statistic* (e.g., mean, sum), denoted $W(\boldsymbol{X})$, is required.[21] As simulating $W(\boldsymbol{X})$ directly can decrease the computational burden of the study data simulation, in some simple settings, we discuss methods for generating $W(\boldsymbol{X})$ directly. However, for many studies (e.g., those collecting censored survival data), it will not be possible to simulate $W(\boldsymbol{X})$ directly, and we will only discuss the individual-level simulation method.

Simulating Study Outcomes Using Parametric Distributions

Plausible study data can be generated by specifying a *parametric* data-generating process $p(\boldsymbol{X}|\boldsymbol{\theta})$. The exact parametric data-generating process will change depending on the proposed study design as it must reflect which model parameters the study will inform and what data should be collected to update these parameters. For example, a randomized controlled trial can be proposed to inform the log odds ratio of a given health event between the current standard and novel treatment while a cohort study would inform the baseline event rate, and a study analyzing administrative claims data would inform costs. Studies can also be proposed to updated multiple model parameters, and the parametric data-generating process can be specified in an arbitrarily complex manner to design increasingly realistic studies.

Irrespective of the complexity of $p(\boldsymbol{X}|\boldsymbol{\theta})$, plausible datasets can be generated from $p(\boldsymbol{X})$ by first simulating from the marginal distribution of the parameters $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta})$ and then simulating from the sampling distribution of the data based on the sampled parameter values $\boldsymbol{x} \sim p(\boldsymbol{X}|\boldsymbol{\theta}^*)$. This generates samples from the *joint* distribution of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ as $p(\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. By generating samples from the joint distribution of $p(\boldsymbol{X}, \boldsymbol{\theta})$ and "ignoring" the samples of $\boldsymbol{\theta}$, we generate datasets from the distribution of the data, $\boldsymbol{x} \sim p(\boldsymbol{X})$, that include both first-order (i.e., individual-level) uncertainty and second-order (i.e., parametric) uncertainty.

Table 1 Representation of a Probabilistic Analysis (PA) Sample with $\mathcal{S}$ Samples for a Set of $\mathcal{P}$ Parameters and $\mathcal{D}$ Decision Options[a]

Probabilistic Analysis Sample								
Parameters			**Net Benefits**			**Simulated Datasets**		
$\theta_1^{(1)}$	$\cdots$	$\theta_{\mathcal{P}}^{(1)}$	$NB_1^{(1)}$	$\cdots$	$NB_{\mathcal{D}}^{(1)}$	$x_1^{(1)}$	$\cdots$	$x_{\mathcal{O}\times\mathcal{M}}^{(1)}$
$\theta_1^{(2)}$	$\cdots$	$\theta_{\mathcal{P}}^{(2)}$	$NB_1^{(2)}$	$\cdots$	$NB_{\mathcal{D}}^{(2)}$	$x_1^{(2)}$	$\cdots$	$x_{\mathcal{O}\times\mathcal{M}}^{(2)}$
$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\theta_1^{(\mathcal{S})}$	$\cdots$	$\theta_{\mathcal{P}}^{(\mathcal{S})}$	$NB_1^{(\mathcal{S})}$	$\cdots$	$NB_{\mathcal{D}}^{(\mathcal{S})}$	$x_1^{(\mathcal{S})}$	$\cdots$	$x_{\mathcal{O}\times\mathcal{M}}^{(\mathcal{S})}$

[a]The bracketed superscript indexes the parameter samples, corresponding net benefits, and simulated datasets.

In practice, $\mathcal{S}$ samples of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta})$ are required in PA and are thus available as part of standard cost-effectiveness analyses that compute the net benefit for each decision option $d = 1, \ldots, \mathcal{D}$.[8] To present the data-generating algorithm, the first 2 columns of Table 1 represent this standard PA, where the parameter samples and net benefits are indexed with a bracketed superscript.

We assume that our study aims to record $\mathcal{O}$ quantities (study outcomes) on $\mathcal{M}$ participants, resulting in $\mathcal{O}\times\mathcal{M}$ measurements in the study. For example, a study could recruit 100 people ($\mathcal{M} = 100$) to measure their blood pressure and quality of life ($\mathcal{O} = 2$). Thus, a single study dataset is denoted as the vector $\boldsymbol{x} = (x_{1,1}, \ldots, x_{\mathcal{O}\times\mathcal{M}})$. The third column of Table 1 demonstrates that each PA parameter sample $\boldsymbol{\theta}^{(s)}$ is used to sample from the conditional distribution of the data, $\boldsymbol{x}^{(s)} \sim p(\boldsymbol{X}|\boldsymbol{\theta}^{(s)})$, to generate the samples $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(\mathcal{S})}$ that follow the marginal distribution of the data $p(\boldsymbol{X})$. We can also consider studies (e.g., cohort or registry studies) that propose collecting the $\mathcal{O}$ individual-level quantities at $\mathcal{T}$ different time points. Again, these studies can be generated using the same algorithm, but each simulated dataset will contain $\mathcal{O}\times\mathcal{M}\times\mathcal{T}$ measurements.

Univariate Data Simulation for Complete Datasets

Initially, we consider studies that collect a single outcome at a single time point for each participant (i.e., $\mathcal{O} = 1$).

Generating binary outcome data. Assume that our decision-analytic model has a parameter, θ_1, that is the proportion of individuals in a population who experience an event (e.g., a stroke) under the current standard treatment. Our current knowledge about this proportion is represented by a prior distribution $p(\theta_1)$, informed from a previous study or a literature search.[30] In our PA, we have $\mathcal{S}$ samples $\{\theta_1^{(1)}, \ldots, \theta_1^{(\mathcal{S})}\}$ drawn from $p(\theta_1)$. Information about θ_1 could be updated by extracting $\mathcal{M}$ individuals from a patient registry and determining whether each individual has experienced the event, resulting in a binary outcome (event v. no event) that can be simulated from a Bernoulli distribution with parameter p equal to the probability of an adverse event. To generate $\mathcal{S}$ datasets from $p(\boldsymbol{X})$, we take each value of $\theta_1^{(s)}$ for $s = 1, \ldots, \mathcal{S}$, and sample $\mathcal{M}$ binary outcomes with parameter $p = \theta_1^{(s)}$. Assuming $\mathcal{S} = 1000$ and $\mathcal{M} = 100$, we can generate this dataset in R as follows:

```
S <- 1000 # Number of simulated datasets
M <- 100 # Number of individuals extracted from the registry
x <- matrix(NA, nrow = S, ncol = M) # Set up empty matrix
theta_1 <- runif(S, 0.1, 0.2) # Distribution for theta_1
for (s in 1:S) # Simulate s = 1,...,S studies
  p <- theta_1[s] # Set the Bernoulli parameter to the s-th
  # value of theta_1
  x[s, ] <- rbinom(n = M, size = 1, prob = p) # Sample M binary
  # event outcomes
}
```

Alternatively, the number of events in each simulated study (i.e., a summary of the study data) can be sampled from a *binomial* distribution with parameter p and the number of "trials" (`size`) equal to $\mathcal{M}$. This highlights the distinction between simulating individual-level data, $\boldsymbol{x}$, and simulating a summary statistic of the individual-level data, $W(\boldsymbol{x})$. This summary statistic is generated in R as follows:

```
M <- 100
Wx <- numeric(length = S) # Set up empty vector
for (s in 1:S) { # Simulate s = 1,...,S studies
  p <- theta_1[s] # Set the Binomial parameter to the s-th
  # value of theta_1
  Wx[s] <- rbinom(n = 1, size = M, prob = p) # Sample count of
  # the event outcomes
}
```

In this example, simulating the data summary is relatively simple and therefore recommended. However, if multiple outcomes will be simulated for each individual (see the multivariate data simulation section), then the individual-level binary outcomes will likely be required.

Generating normally distributed continuous data. Assume that the decision-analytic model has a parameter, θ_2, that represents the mean systolic blood pressure in the population. The current prior uncertainty about θ_2, obtained through a previous study on θ_2, is modeled in $p(\theta_2)$. Additional information could be gathered in a cross-sectional study that measures the blood pressure in $\mathcal{M}$ individuals. We assume that the individual-level systolic blood pressure follows a normal distribution from which we can simulate a dataset for $\mathcal{M}$ study participants. To generate $\mathcal{S}$ datasets from the marginal distribution of the data, we take each value of $\theta_2^{(s)}$ for $s = 1, \ldots, \mathcal{S}$ and sample from a normal distribution with mean $\mu = \theta_2^{(s)}$. The variance for the normal distribution represents the *individual-level* variance in blood pressure and can either be assumed known or assigned a probability distribution that represents our uncertainty in the individual-level variance of the systolic blood pressure. Crucially, this individual-level variance, which can be extracted from the literature or estimated from available individual-level data, is unlikely to be equal to the variance of θ_2, which represents the uncertainty in our knowledge about the parameter. Note that an estimate of the individual-level variance is required for standard sample size calculations, used to ensure that a hypothesis test undertaken with the trial data has sufficient power.[31] Assuming $\mathcal{S} = 1000$, $\mathcal{M} = 100$, and an individual-level variance (v) of 80, these data are simulated in R as follows:

```
S <- 1000
M <- 100;
x <- matrix(nrow = S, ncol = M) # Set up empty matrix
theta_2 <- runif(S, 120, 130) # Hypothetical distribution
  # for theta_2
v <- 80
for (s in 1:S) { # Simulate s = 1,...,S studies
   mu <- theta_2[s] # Set the Normal mean parameter to the
   # s-th value of theta_2
   x[s, ] <- rnorm(n = M, mean = mu, sd = sqrt(v)) # Sample M
   # blood pressure measures
}
```

Alternatively, if the study is aiming to estimate the mean systolic blood pressure, then the summary statistic $W(\boldsymbol{x})$ (i.e., the study mean systolic blood pressure) can be simulated directly from the sampling distribution of the mean. In this case, the study-level mean blood pressure would be simulated from a normal distribution with mean $\mu = \theta_2^{(s)}$ and standard deviation equal to the square root of the individual-level variance divided by the sample size $\mathcal{M}$ (i.e., the standard error of the mean). R code for this simulation is given as follows:

```
M <- 100
v <- 80
Wx <- numeric(length = S) # Set up empty vector
for (s in 1:S) { # Simulate s = 1,...,S studies
  mu <- theta_2[s] # Set the Normal mean parameter to the s-th
  # value of theta_2
  Wx[s] <- rnorm(n = 1, mean = mu, sd = sqrt(v / M)) # Sample
  # study mean BP
}
```

Many summary statistics are approximately normal (e.g., the log odds ratio or log hazard ratio), allowing us to potentially adapt this simulation method for other summary statistics. However, the standard error for these alternative summary statistics must be specified correctly, which can be challenging especially when considering variable sample sizes for the study. Thus, it may be more appropriate to generate individual-level data and then calculate the summary statistic from the simulated dataset by analyzing the simulated data as if it were collected during a study (see the data on relative effectiveness section below).

Generating time-to-event data. Assume that our decision-analytic model has a parameter, θ_3, that represents the probability that a patient's cancer progresses within a 1-month period on the current standard treatment. The prior distribution of this transition probability, potentially estimated from the control arm in a clinical trial or from administrative data, is represented by $p(\theta_3)$ and will be updated by measuring the time to cancer progression in $\mathcal{M}$ individuals from a cancer registry. Assuming that the rate of progression is constant over time, we can simulate time-to-progression data from an exponential distribution with rate, $r = -\log(1 - \theta_3)$. Thus, generating $\mathcal{S}$ datasets takes each value of $\theta_3^{(s)}$ for $s = 1, \ldots, \mathcal{S}$ and samples $\mathcal{M}$ time-to-progression data from an exponential distribution with parameter $r = -\log(1 - \theta_3^{(s)})$. Assuming $\mathcal{S} = 1000$ and $\mathcal{M} = 100$, the following R code generates the following data:

```
S <- 1000; theta_3 <- runif(S, 0.2, 0.3) # Hypothetical
  # distribution for theta_3
M <- 100
x <- matrix(nrow = S, ncol = M) # Set up empty matrix
for (s in 1:S) { # Simulate s = 1,...,S studies
  r <- -log(1 - theta_3[s]) # Derive rate from s-th value of
  # the transition probability
  x[s, ] <- rexp(n = M, rate = r) # Sample M times-to-
  # progression
}
```

Alternative time-to-event distributions are also available (e.g., Weibull, Gamma) but have different parameterizations of the data-generating process. These distributions are more complex because they also have more than 1 parameter. Assume that our decision-analytic model is a partitioned survival model with a Weibull distribution estimating progression-free survival times for the current standard treatment and parameterized in terms of θ_4 and θ_5. Uncertainty in (θ_4, θ_5) is represented by the joint distribution $p(\theta_4, \theta_5)$ and will be updated by a study that collects time-to-progression data for $\mathcal{M}$ individuals. To generate $\mathcal{S}$ datasets, we take each pair of values $\theta_4^{(s)}, \theta_5^{(s)}$ for $s = 1, \ldots, \mathcal{S}$ and sample $\mathcal{M}$ time-to-progression data from a Weibull distribution with correlated parameters $\theta_4^{(s)}, \theta_5^{(s)}$.[32] Assuming $\mathcal{S} = 1000$ and $\mathcal{M} = 100$, R code for this is as follows:

```
S <- 1000
# Correlated joint distribution for theta_4 and theta_5
# (Column 1: theta_4, Column 2: theta_5)
theta_4_5 <- MASS::mvrnorm(S,
        c(5,6),
        matrix(c(0.3, 0.1, 0.1, 0.5), nrow = 2))
M <- 100
x <- matrix(nrow = S, ncol = M) # Set up empty matrix
for (s in 1:S) { # Simulate s = 1,...,S studies
   shape <- theta_4_5[s, 1] # Weibull shape parameter from
   # s-th value of theta_4
   scale <- theta_4_5[s, 2] # Weibull scale parameter from
   # s-th value of theta_5
   x[s, ] <- rweibull(n = M, shape = shape, scale = scale)
   # Sample M times-to-progression
 }
```

Note that choosing the appropriate individual-level distribution for this data simulation can be challenging, and methods are currently being developed to adapt the EVSI calculation method itself when the survival distribution is unknown.[33] However, these methods still need to simulate from a range of survival distributions and will thus require the methods presented here.

Generating utility data. Next, assume that our health economic model has a parameter, θ_6, that represents the mean utility for a specific health state (e.g., the preprogression state). Information about θ_6 could arise from a previous utility elicitation exercise and is encoded in a beta prior distribution $p(\theta_6)$. Additional information on the utility could be gathered through a utility elicitation study among individuals in the given health state (e.g., through the use of a standard gamble method). We can assume that this utility score follows a beta distribution with a mean of θ_6 and an individual-level variance v obtained from a previous study. To simulate these data, the mean and variance must be translated into the parameters of the beta distribution, which we achieve using the function `calculate_beta_parameters` below. The following code generates $\mathcal{S} = 1000$ datasets for a study collecting utility scores from $\mathcal{M} = 100$ individuals:

```
S <- 1000;theta_6 <- rbeta(S, 70, 15) # Hypothetical
  # distribution for theta_6
M <- 100
v <- 0.04
x <- matrix(nrow = S, ncol = M) # Set up empty matrix
calculate_beta_parameters <- function(mean, sd){
  # Function to estimate beta parameters from mean and
  # standard deviation
  shape1 <- ((1 - mean) / sd ^ 2 - 1 / mean) * mean ^ 2
  shape2 <- shape1 * (1 / mean - 1)
  # Return the calculated parameters.
  return(list(shape1 = shape1,
      shape2 = shape2))
}
for (s in 1:S) { # Simulate s = 1,...,S studies
  # Derive beta parameters with iteration specific mean
  params <- calculate_beta_parameters(theta_6[s], sqrt(v))
  x[s, ] <- rbeta(n = M, shape1 = params$shape1,
      shape2 = params$shape2) # Sample M times-to-progression
}
```

There are a large range of study types (e.g., those that collect data on costs or resource use) that we are not able to address directly in this tutorial. However, the general-purpose algorithm can be adapted to simulate from the relevant distributions (e.g., log-normal distribution for costs).[1]

Multivariate Data Simulation for Complete Datasets

If the proposed study collects more than 1 outcome for each study participant, $\mathcal{O}>1$, and/or outcomes at more

than 1 time point, alternative methods will be required. In this framework, any study where the individuals receive different interventions (e.g., randomized controlled trials) are defined as multivariate data collection exercises. This is because we specify the treatment that the individual receives as one of the $\mathcal{O}$ quantities of interest. Thus, $\mathcal{O}>1$ as we record the treatment and at least 1 outcome, demonstrated in the data on relative effectiveness section below.

Independent multivariate data simulation. If the quantities generated for each participant are assumed to be independent, conditional on $\boldsymbol{\theta}$, a separate univariate data-generating process can be specified for each of the $\mathcal{O}$ quantities of interest and then combined into a single dataset. Assuming the data are independent conditional on the parameters does not mean that the data are uncorrelated as any correlations in the model parameters, embodied in $\boldsymbol{\theta}$, would generate correlated patient-level study data. A combined study that investigates $\mathcal{M} = 100$ participants and records whether they experience an adverse event and their times to progression can be generated in `R` as follows:

```
S <- 1000
O <- 2
M <- 100
x <- array(dim = c(M, O, S)) # Set up empty array
for (s in 1:S) { # Simulate s = 1,...,S studies
  p <- theta_1[s] # Set the Bernoulli parameter to the
  # s-th value of theta_1
  r <- -log(1 - theta_3[s]) # Derive rate from s-th value of
  # the transition probability
  x[ , 1, s] <- rbinom(n = M, size = 1, prob = p) # Sample M
  # binary adverse outcomes
  x[ , 2, s] <- rexp(n = M, rate = r) # Sample M times-to-
  # progression
}
```

This code does not store the data using the spreadsheet structure demonstrated in Table 1, but it uses a 3-dimensional array with $\mathcal{M}$ rows for each study participant, $\mathcal{O}$ columns for each recorded quantity, and $\mathcal{S}$ matrix *slices* (the third dimension) for each simulation. This structure makes it easier to analyze data separately for each simulation if this is required to estimate the summary statistics.

Dependent multivariate data simulation. Multivariate data simulation is more complex when the simulated quantities are correlated for each participant (e.g., if participants with shorter survival times are more likely to experience adverse events). This correlation must be specified when we generate multivariate data and can either be assumed fixed or assigned a probability distribution that represents our uncertainty about the correlation. If we ignore the correlation, we are implicitly assuming that it is zero, with certainty. Thus, even if evidence about the correlation structure is lacking, it is important to assess whether this assumption of zero correlation is valid. In general, the correlation can be informed 1) by the literature, although reporting on correlation is often lacking, and you may need to request this information from the authors; 2) by calculating the correlation in available data; or 3) through expert elicitation.[34]

One method to generate correlated data initially generates uncorrelated data and then reorders the simulated dataset to achieve the required correlation.[35,36] These reordering methods are implemented in the `R` function `postSimOpt`, which generates correlated data with a given correlation matrix.[37] If we are generating correlated data similar to the previous example recording adverse events and time-to-progression data from $\mathcal{M} = 100$ participants, then we can reorder the data from the previous example to have a correlation of –0.2 using `R` as follows:

```
library(SimJoint) # Package containing function to reorder
  # data
S <- 1000
O <- 2
M <- 100
correlation <- matrix(c(1, -0.2, -0.2, 1), nrow = 2)
# Specify the correlation matrix
x <- array(dim = c(M, O, S)) # Set up empty array
for (s in 1:S) { # Simulate s = 1,...,S studies
  p <- theta_1[s] # Set the Bernoulli parameter to the s-th
  # value of theta_1
  r <- -log(1 - theta_3[s]) # Derive rate from s-th value of
  # the transition probability
  x[ , 1, s] <- rbinom(n = M, size = 1, prob = p) # Sample M
  # binary adverse outcomes
  x[ , 2, s] <- rexp(n = M, rate = r) # Sample M times-to-
  # progression
  # Reorder the columns so they are correlated
  x[ , , s] <- postSimOpt(x[, , s], correlation) $X
}
```

Correlated data can also be generated using regression to specify the dependencies between the quantities of interest. The regression method decomposes the joint distribution of these quantities into conditional and

marginal distributions, where the conditional distributions are defined using regression models. This method can generate data for $\mathcal{O}$ correlated quantities of interest, X_o, $o = 1, ...\mathcal{O}$ by initially generating a value of X_1 from its marginal distribution, before proceeding to generate X_2 conditional on X_1, with the relationship specified using regression. Following this, X_3 can be generated based on X_1 and X_2 and so on. If $\mathcal{O}$ is small, then the required regression models may have been published, but as the number of outcomes increases, IPD will be required to fit these models. The data generation should consider uncertainty in the parameters of the regression model, specified either by fitting the regression models using Bayesian methods or sampling the regression coefficients from their sampling distribution. This sampling distribution is approximately multivariate normal with the variance-covariance matrix estimated when the regression models are fit in standard software. Thus, if published regression models are used, the variance of the regression parameters must also be extracted. Using the previous example and assuming that its first simulated dataset is actually IPD recording adverse events and time-to-progression data that are saved in a data frame called dat, the following code generates correlated data using the regression method:

```
library(MASS) # Package to simulate from multivariate normal
  # distribution
S <- 1000
M <- 100; O <- 2
dat <- as.data.frame(x[ , , 1])
# Generalised Linear Model to predict adverse event
  # probability from times-to-progression
mod <- glm(AE Time_Prog, data = dat, family = ''binomial'')
theta_reg <- mvrnorm(S, coef(mod), vcov(mod)) # Sampling
  # distribution of coefficients

x <- array(dim = c(M, O, S)) # Set up empty array
for (s in 1:S) { # Simulate s = 1,...,S studies
   r <- -log(1 - theta_3[s]) # Derive rate from s-th value of
   # the transition probability
   x[ , 2, s] <- rexp(n = M, rate = r) # Sample M times-to-
   # progression
   mod$coefficients <- theta_reg[s, ] # Set the coefficients
   # to their s-th value
   # Predict probability of an adverse event from the simulated
   # times-to-progression
   p.ind <- predict(mod, data.frame(Time_Prog = x[, 2, s]),
  type = ''response'')
   x[ , 1, s] <- rbinom(n = M, size = 1, prob = p.ind) # Sample M
   # binary adverse outcomes
}
```

These methods can be combined with the uncorrelated data generation processes to generate both dependent and independent data for the proposed study.

Data on relative effectiveness. Data from a proposed randomized control trial, which updates uncertainty in the log odds ratio of an event on a novel intervention compared to the current standard treatment (θ_7), also require correlated multivariate data generation. The first quantity of interest is an indicator $\mathbb{I}$, highlighting which treatment each participant receives. In an equally randomized 2-arm trial, this is generated from a Bernoulli distribution with probability 0.5, with a 1 representing that the participant has been randomized to receive the novel intervention. To calculate the patient-level probability of experiencing the outcome event of interest from this indicator, we must combine the s th simulated values of $\theta_7^{(s)}$ with the simulated values of the baseline probability of experiencing the event under the standard treatment, denoted $\theta_8^{(s)}$. (Note that information on the baseline probability of the event can, and often should, come from a different source than the information to inform θ_8, i.e., the baseline event rate comes from administrative data, while a previous clinical trial would inform the relative effectiveness.) The individual-level log odds of experiencing the event can then be computed by adding $\theta_7^{(s)} \times \mathbb{I}$ to $logit(\theta_8^{(s)})$. The individual-level probability of the event is then calculated from $logit^{-1}\{logit(\theta_7^{(s)}) + \theta_8^{(s)} \times \mathbb{I}\}$, and the individual-level response can be generated from a Bernoulli distribution with these probabilities. The summary statistic (e.g., the observed log odds ratio) can then be estimated by fitting a generalized linear model to the s th dataset as though the simulated data were observed. The following R code implements this method for a study collecting data on $\mathcal{M} = 100$ participants:

```
library(boot) # Package for logit and inv.logit
S <- 1000
M <- 100; O <- 2
theta_7 <- rnorm(S, 1.2, 0.1) # Hypothetical distribution
  # for log odds ratio
theta_8 <- runif(S, 0.2, 0.3) # Hypothetical distribution
  # for baseline risk
x <- array(dim = c(M, O, S)) # Set up empty array
Wx <- numeric(length = S) # Set up empty vector for simulated
   # summary statistic
for (s in 1:S) { # Simulate s = 1,...,S studies
    # Sample M treatment indicators
    x[ , 1, s] <- rbinom(n = M, size = 1, p = 0.5)
```

(continued)

```
    # Calculate s-th baseline log odds
    baseline.logodds <- logit(theta_8[s])
    # Calculate odds for treated group from baseline log odds
    # and the s-th log odds ratio
    individual.logodds <- baseline.logodds + theta_7[s] *
    x[ , 1, s]
    # Calculate probability from log odds
    individual.prob <- inv.logit(individual.logodds)
    # Sample M binary outcomes
    x[ , 2, s] <- rbinom(n = M, size = 1, prob =
    individual.prob)
    # Create a dataframe with the data
    data.complete <- data.frame(x[, , s])
    names(data.complete) <- c(''Treatment,'' ''Outcome'')
    # Generalised linear model to compute odds ratio for the
    s-th dataset
    Wx[s] <- glm(Outcome  Treatment, data = data.complete,
    family = ''binomial'')$coef[2]
}
```

This example uses binary outcomes and log odds ratios as a measure of relative effect. If an alternative outcome type and/or measure of relative effect is used, then this method must be adapted to translate the parameters to the additive scale and back to generate the data. We provide code to implement this method for survival outcomes and log hazard ratios in the supplementary material.

Finally, there are many methods for generating correlated data that are not discussed in this tutorial. Copulas are a class of statistical models that combine univariate marginal distributions and a multivariate correlation structure and can generate correlated data.[38] Elsewhere, methods can ensure that simulated data preserve their rank (i.e., in situations where 1 outcome must be larger than another).[39] Microsimulation models or discrete-event simulations can also generate interrelated individual event data in a highly flexible but more computationally intensive manner.[40,41]

Realistic Study Designs

Realistic studies can encounter issues with missing values, loss to follow-up, and censoring, which should be included in our data simulation procedure.[6]

Missingness. Data that are not recorded during a study (i.e., missing data) are commonly accounted for in study design and analysis.[42] Thus, simulating missing values based on knowledge about the potential rate of missingness will often be required. A "missingness indicator" equals 1 if the participant's data are missing and 0 otherwise. This can be used to simulate missingness using a Bernoulli distribution with the probability equal to the expected level of missingness, obtained from the literature or expert opinion. Once the missingness indicator has been generated, participants with a missingness indicator of 1 are then "deleted" from the simulated dataset. If the study collects multivariate outcomes, then missingness can be considered separately for each outcome. The simplest type of missingness (i.e., missing completely at random) generates the missingness indicator independent of the quantities of interest[43] with an example assuming 10% missing data given as follows:

```
S <-1000; theta_2 <-runif(S, 120, 130) # Hypothetical
  # distribution for theta_2
M <-100; v <-80
x <-matrix(nrow = S, ncol = M) # Set up empty matrices
for (s in 1:S) { # Simulate s = 1,...,S studies
   mu <-theta_2[s] # Set the Normal mean parameter to the s-th
   # value of theta_2
   x[s, ] <-rnorm(n = M, mean = mu, sd = sqrt(v)) # Sample M
   # blood pressure measures
   missing <-rbinom(n = M, size = 1, prob = 0.1) # Sample
   # missingness indicator
   x[s, which(missing == 1)] <-NA # Knock out the missing
   # observations
}
```

A correlation between the data and the missingness indicator (i.e., where participant outcomes or traits lead to higher levels of missingness) can also be assumed and would induce bias in estimates from the data and EVSI if it is not accounted for properly. If this type of missingness is used, then the method for updating the distribution of the model parameters, based on the data, would also need to be adjusted using common methods for addressing missing data.[42]

Censoring in time-to-event data. Censoring is commonly encountered when working with time-to-event data; for example, right-censored data include the information that a participant did not experience an event during the study but do not record when (or if) the event is experienced after the study's observation period ended. Censoring is modeled by adding a "censoring indicator" to the dataset, which equals 0 if the data point is censored and 1 if it is not. To generate censored survival data, we first generate the event time for each participant from a suitable uncensored model (cf. generating time-to-event data). We then generate a potential "censoring time" for each participant; this can either be a fixed number (i.e.,

Table 2 Representation of the Bootstrap Estimation Method for the Parameter θ_8 Based on an Initial Sample of Size N

Simulation	y_1	y_2	y_3	...	y_N	θ_8
1	$y_1^{(1)}$	$y_2^{(1)}$	$y_3^{(1)}$	...	$y_N^{(1)}$	$\theta_8^{(1)}$
2	$y_1^{(2)}$	$y_2^{(2)}$	$y_3^{(2)}$	...	$y_N^{(2)}$	$\theta_8^{(2)}$
⋮	⋮	⋮	⋮	⋱	⋮	⋮
$\mathcal{S}$	$y_1^{(\mathcal{S})}$	$y_2^{(\mathcal{S})}$	$y_3^{(\mathcal{S})}$	...	$y_N^{(\mathcal{S})}$	$\theta_8^{(\mathcal{S})}$

all patients are censored at the end of the study follow-up) or simulated from a different time-to-event distribution with parameters estimated to reflect patterns of dropout or loss to follow-up seen in similar studies.[44] If the censoring event occurs before the event, we change the event time to the censoring time and the censoring indicator to 0. An example where time-to-progression data are censored at 6 months is given as follows:

```
S <-1000; theta_3 <-runif(S, 0.2, 0.3) # Hypothetical
  # distribution for theta_3
M <-100
x <-matrix(nrow = S, ncol = M) # Set up empty matrix
censoring_time <-6
for (s in 1:S) { # Simulate s = 1,...,S studies
  r <- -log(1 - theta_3[s]) # Derive rate from s-th value of
  # the transition probability
  x[s, ] <- rexp(n = M, rate = r) # Sample M times-to-
  # progression
}
censoring_indicator <- (x > censoring_time) # Set indicator
  # for times > 6 months
x[censoring_indicator] <- censoring_time # Set censored
  # times to 6 months
```

This code implements right-censoring, commonly seen in randomized control trials, but a similar method could simulate left-censored data, where the event time is not observed if it occurs before the censoring time. Finally, interval censoring, where only the time interval in which the event occurs is known, requires a more complex specification.

Simulating Study Outcomes Using Nonparametric Resampling

If the decision-analytic model is based on IPD, we could investigate whether there is value in collecting additional data with the same (or a similar) study design. Given IPD are available, we could generate data in this setting by resampling the IPD and avoid specifying parametric distributions for the data. Resampling from IPD, denoted $\boldsymbol{y}$, can characterize parameter uncertainty using bootstrap methods,[45] but these methods must be extended to generate the range of plausible datasets from $p(\boldsymbol{X})$. Assume that a parameter for a decision-analytic model, θ_8, can be estimated as a function of the IPD, $\theta_8 = H(\boldsymbol{y})$. The uncertainty in θ_8 can be estimated by resampling $\mathcal{S}$ times from $\boldsymbol{y}$ with replacement to create multiple pseudo-datasets $\boldsymbol{y}^{(s)}$, $s = 1, ..., \mathcal{S}$ before estimating the model parameter $\theta_8^{(s)} = H\left(\boldsymbol{y}^{(s)}\right)$ (Table 2).

To simulate a dataset from $p(\boldsymbol{X})$ with $\mathcal{M}$ participants for each row of the PA dataset, we should resample $\mathcal{M}$ values with replacement from each dataset $\boldsymbol{y}^{(s)}$, $s = 1, ..., S$ (i.e., resample from each row of Table 2). This is equivalent to generating the data from $p(\boldsymbol{X}|\theta_8^{(s)})$. The following displays the R code for this resampling algorithm:

```
S <- 1000
N <- 150; M <- 100
y <- runif(N, 10, 30) # Hypothetical IPD
x <- matrix(nrow = S, ncol = M) # Set up empty matrix
for (s in 1:S) { # Simulate s = 1,...,S studies
   y_s <- sample(y, N, replace = TRUE) # Bootstrap sample from y
   x[s, ] <- sample(y_s, M, replace = TRUE) # Sample M IPD values
   # from y_s
}
```

This resampling method can also generate datasets that are similar to the IPD. For example, if the proposed study targets younger participants than the previous study, we could perform a weighted resampling to sample the younger patients more frequently. We could also sample a subset of the quantities from the previous study to evaluate the value of a more targeted study or plan a study with a shorter follow-up.

Once we have generated our resampled datasets, the efficient EVSI estimation procedures require different adaptions to estimate EVSI. Methods that require

Bayesian updating (e.g., the standard Monte Carlo method and the moment matching method)[46] must use an adapted bootstrap algorithm, which we are currently developing, to approximate the Bayesian updating without specifying $p(\boldsymbol{\theta})$ and $p(\boldsymbol{X}|\boldsymbol{\theta})$ analytically. Methods that require a summary statistic (e.g., the regression-based method)[16] can be used by calculating the parameter using the function $H(\cdot)$ for each simulated dataset. Note that one of the EVSI calculation methods is based on evaluating the likelihood function of the data and so cannot be used with this resampling method.[19]

Discussion

EVSI can be used to optimize study designs to generate data to support decision making in HTA processes, which are often based on decision-analytic models.[47] EVSI can formalize the decision to collect additional information before making policy decisions in health, thereby ensuring that effective and efficient treatments are available to patients.[48–50] This tutorial supports the increased use of EVSI by researchers, decision makers, and industry partners by presenting a range of methods to generate simulated datasets for EVSI calculation.

Recent research has allowed practical EVSI calculations through the development of efficient estimation methods,[21] which generally require simulated datasets from a proposed future study. The methods presented in this tutorial can be used to simulate datasets from randomized trials and observational studies with a range of outcome types, including uni- and multivariate datasets. Furthermore, they support the modeling of imperfect study conduct and incomplete data collection. Finally, they are applicable with and without individual patient-level data. We demonstrate these methods using `R` code and, where appropriate, with `Excel` spreadsheets included in the supplementary material. Once we have simulated the datasets from the proposed study, the final computation of EVSI depends on the selected algorithm, as detailed in Kunst et al.[21]

Accurate EVSI estimation requires realistic data simulation.[6] These datasets should reflect our judgments about the data, encoded in our chosen parameter distributions $p(\boldsymbol{\theta})$ and data-generating process. Thus, they do not need to reflect a dataset that has previously been collected, making it challenging to determine if the simulated datasets are "correct." However, when developing the simulation method, biological plausibility can and should be checked (e.g., determine that all simulated survival times are within the life span of a human). It may also be worthwhile to check whether the simulated data reflect the specified inputs (e.g., calculate the individual-level variance for each simulation and check if it is approximately equal to the specified variance). As the number of simulated datasets is large, these checks may only be possible for a small number of the datasets and can be used for validation.

As studies can be designed with almost infinite complexities, many study designs that are relevant to health economic decision making could not be included in this tutorial. For example, simulating data on utilities is potentially more complex than the method presented in this tutorial as health states are often ranked, and the data simulation should take this into account, potentially through previously developed methods.[39] Recent research has also proposed methods for EVSI calculation when the survival distribution is unknown and may change based on the future data.[33] Furthermore, studies based on long-term longitudinal cohorts will require complex multivariate data generation and missing data patterns. Finally, the estimation of study costs to compute ENBS and optimize study design has received limited discussion in the literature[3] despite its importance to ensure accurate research prioritization.

Conclusion

This tutorial presents a general-purpose algorithm for generating simulated datasets from a probabilistic analysis and explored common correlated and uncorrelated data types. This method is demonstrated in several examples but can be extended to more complex study designs, as required. Hence, this tutorial facilitates practical EVSI calculations and allows research design and prioritization based on ENBS.

Acknowledgments

The authors thank the Collaborative Network on Value of Information for their comments and discussion on this manuscript. In particular, the authors thank Ed Wilson, Christopher Jackson, and Fernando Alarid-Escudero for their comments on earlier versions of this manuscript.

ORCID iDs

Anna Heath https://orcid.org/0000-0002-7263-4251
Mark Strong https://orcid.org/0000-0003-1486-8233
David Glynn https://orcid.org/0000-0002-0989-1984

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at http://journals.sagepub.com/home/mdm.

References

1. Briggs A, Sculpher M, Claxton K. *Decision Modelling for Health Economic Evaluation*. Oxford, UK: Oxford University Press; 2006.
2. Raiffa H, Schlaifer R. *Applied Statistical Decision Theory*. Boston, MA: Division of Research, Graduate School of Business Administration, Harvard University; 1961.
3. Conti S, Claxton K. Dimensions of design space: a decision-theoretic approach to optimal research design. *Med Decis Making*. 2009;29(6):643–60.
4. Willan AR, Eckermann S. Optimal clinical trial design using value of information methods with imperfect implementation. *Health Econ*. 2010;19(5):549–61.
5. Welton NJ, Thom HHZ. Value of information: we've got speed, what more do we need? *Med Decis Making*. 2015; 35(5):564–6.
6. Rothery C, Strong M, Koffijberg HE, et al. Value of information analytical methods: report 2 of the ISPOR Value of information analysis emerging good practices task force. *Value Health*. 2020;23(3):277–86.
7. Ades AE, Lu G, Claxton K. Expected value of sample information calculations in medical decision modeling. *Med Decis Making*. 2004;24(2):207–27.
8. Claxton K, Sculpher M, McCabe C, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Econ*. 2005;14:339–47.
9. Canadian Agency for Drugs and Technologies in Health. *Guidelines for the Economic Evaluation of Health Technologies*. 4th ed. Ottawa, ON: CADTH; 2017.
10. Department of Health and Ageing. *Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee: Version 4.3*. Canberra, Australia: Department of Health: Australian Government; 2008.
11. EUnetHTA. *Methods for Health Economic Evaluations: A Guideline Based on Current Practices in Europe: Second Draft*. Rotterdam, Netherlands: EUnetHTA; 2014.
12. Agency NM. Guidelines for the submission of documentation for single technology assessment (STA) of pharmaceuticals. 2018. Available from: https://legemiddelverket.no/english/public-funding-and-pricing/documentation-for-sta/
13. Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. *Value Health*. 2012;15(6):835–42.
14. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment. *Pharmacoeconomics*. 2006; 24(4):355–71.
15. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in healthcare evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health*. 2003;6(1):9–17.
16. Strong M, Oakley JE, Brennan A, Breeze P. Estimating the expected value of sample information using the probabilistic sensitivity analysis sample: a fast, nonparametric regression-based method. *Med Decis Making*. 2015;35(5): 570–83.
17. Heath A, Manolopoulou I, Baio G. Estimating the expected value of sample information across different sample sizes using moment matching and nonlinear regression. *Med Decis Making*. 2019;39(4):346–58.
18. Jalal H, Alarid-Escudero F. A Gaussian approximation approach for value of information analysis. *Med Decis Making*. 2018 2;38(2):174–88.
19. Menzies NA. An efficient estimator for the expected value of sample information. *Med Decis Making*. 2016 4;36(3):308–20.
20. Heath A, Kunst NR, Jackson C, et al. Calculating the expected value of sample information in practice: considerations from three case studies. Available from: http://arxiv.org/abs/1905.12013
21. Kunst N, Wilson EC, Glynn D, et al. Computing the expected value of sample information efficiently: practical guidance and recommendations for four model-based methods. *Value Health*. 2020;23(6):734–42.
22. Wicklin R. *Simulating Data with SAS*. Cary, NC: SAS Institute; 2013.
23. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279–92.
24. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019; 38(11):2074–102.
25. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation. *Pharmacoeconomics*. 2011;29(1):35–49.
26. Goldhaber-Fiebert JD, Stout NK, Goldie SJ. Empirically evaluating decision-analytic models. *Value Health*. 2010; 13(5):667–74.
27. Alarid-Escudero F, Gulati R, Rutter CM. Validation of microsimulation models used for population health policy. In: Apostolopoulos Y, Lich KH, Lemke MK, eds. *Complex Systems and Population Health: A Primer*. Oxford, UK: Oxford University Press; 2020. p 227–40.
28. Rubin DB. Statistical disclosure limitation. *J Official Stat*. 1993;9(2):461–8.
29. Nowok B, Raab GM, Dibben C, et al. synthpop: bespoke creation of synthetic data in R. *J Stat Softw*. 2016;74(11): 1–26.
30. Briggs AH, Goeree R, Blackhouse G, O'Brien BJ. Probabilistic analysis of cost-effectiveness models: choosing between treatment strategies for gastroesophageal reflux disease. *Med Decis Making*. 2002;22(4):290–308.
31. Chow SC, Shao J, Wang H, Lokhnygina Y. *Sample Size Calculations in Clinical Research*. Boca Raton, FL: CRC Press; 2017.
32. Degeling K, IJzerman MJ, Koopman M, Koffijberg H. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. *BMC Med Res Methodol*. 2017;17(1):1–12.
33. Vervaart M, Aas E, Claxton K, Strong M, Welton NJ, Wisløff T. Expected value of sample information for

survival data from ongoing trials. Paper presented at the 42nd Annual Meeting of the Society for Medical Decision Making October 2021, Virtual conference; 2020.
34. O'Hagan A, Buck CE, Daneshkhah A, et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. New York, NY: John Wiley; 2006.
35. Iman RL, Conover WJ. A distribution-free approach to inducing rank correlation among input variables. *Commun Stat Simul Comp*. 1982;11(3):311–34.
36. Ruscio J, Kaczetow W. Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behav Res*. 2008;43(3):355–81.
37. Liu CW. SimJoint: simulate joint distribution. 2020. R package version 0.3.7. Available from: https://CRAN.R-project.org/package = SimJoint
38. Nelsen RB. *An Introduction to Copulas*. New York, NY: Springer Science & Business Media; 2007.
39. Goldhaber-Fiebert JD, Jalal HJ. Some health states are better than others: using health state rank order to improve probabilistic analyses. *Med Decis Making*. 2016;36(8): 927–40.
40. Krijkamp EM, Alarid-Escudero F, Enns EA, Jalal HJ, Hunink MM, Pechlivanoglou P. Microsimulation modeling for health decision sciences using R: a tutorial. *Med Decis Making*. 2018;38(3):400–22.
41. Caro JJ, Möller J, Getsios D. Discrete event simulation: the preferred technique for health economic evaluations? *Value Health*. 2010;13(8):1056–60.
42. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. New York, NY: John Wiley; 2019.
43. Heitjan DF, Basu S. Distinguishing "missing at random" and "missing completely at random." *Am Stat*. 1996;50(3): 207–13.
44. Royston P. Tools to simulate realistic censored survival-time distributions. *Stata J*. 2012;12(4):639–54.
45. Efron B. Better bootstrap confidence intervals. *J Am Stat Assoc*. 1987;82(397):171–85.
46. Heath A, Manolopoulou I, Baio G. Efficient Monte Carlo estimation of the expected value of sample information using moment matching. *Med Decis Making*. 2018;38(2):163–73.
47. McKenna C, Claxton K. Addressing adoption and research design decisions simultaneously: the role of value of sample information analysis. *Med Decis Making*. 2011;31(6): 853–65.
48. McKenna C, Chalabi Z, Epstein D, Claxton K. Budgetary policies and available actions: a generalisation of decision rules for allocation and research decisions. *J Health Econ*. 2010;29(1):170–81.
49. McKenna C, Soares M, Claxton K, et al. Unifying research and reimbursement decisions: case studies demonstrating the sequence of assessment and judgments required. *Value Health*. 2015;18(6):865–75.
50. Grimm SE, Strong M, Brennan A, Wailoo AJ. The HTA risk analysis chart: visualising the need for and potential value of managed entry agreements in health technology assessment. *Pharmacoeconomics*. 2017;35(12):1287–96.